Text-to-Image: The Future Generation

Abstract

This study conducts qualitative and quantitative comparisons among four state-of-the-art text-to-image generative models: Stable diffusion, DALLE, CLIP, and Kandinsky. The focus is on evaluating image quality, diversity, and alignment with textual descriptions, highlighting the superiority of Controllable Generative AI. Additionally, to tackle the challenge of aligning creativity with functionality, we propose a framework for Controllable Generative AI in Biomimetic Design, integrating constraints and control over the creative process to generate innovative and feasible designs. The code for our implementation is accessible via the link: https://github.com/JohnsonForSure/ML-in-Bioinfo-and-Healthcare.

1. Introduction

Our perception of the world encompasses multiple sensory modalities, including sight, sound, touch, smell, and taste (Baltrušaitis et al., 2018). Multimodal machine learning endeavors to develop models capable of interpreting and associating information from these diverse modalities (Yadav & Vishwakarma, 2020). Among the various multimodal tasks, text-to-image (T2I) synthesis stands out, involving the generation of images corresponding to given textual descriptions. This task holds immense potential across various domains such as image editing, video games, and computer-aided design, necessitating a deep comprehension of the entities being created (Ramesh et al., 2022; Liu et al., 2021).

Recent advancements in Generative Adversarial Networks (GANs), autoregressive models, and diffusion models have propelled the frontier of T2I synthesis (Zameshina et al., 2023). GANs engage in a competitive framework with a generator and discriminator network, striving to produce realistic images indistinguishable from genuine data (Van Looveren et al., 2021). Autoregressive models, conversely, generate data iteratively, conditioning each step on preceding information (Du et al., 2022). These models, based on transformer architectures, utilize self-attention mechanisms to capture dependencies and relationships within input sequences (Esser et al., 2021). On the other hand, diffusion models represent a paradigm shift in machine learning, introducing noise to data and progressively refining it through denoising iterations (Meng et al., 2021).

To evaluate the efficacy of these generative AI models, we conduct qualitative and quantitative comparisons across four state-of-the-art T2I generation paradigms: Stable diffusion (Rombach et al., 2022), DALLE (Ramesh et al., 2021), CLIP (Ramesh et al., 2022), and Kandinsky generative AI model (Razzhigaev et al., 2023). Our assessment examines the quality, diversity, and alignment of generated images with given textual descriptions across various levels, with a focus on highlighting the superior performance of Generative AI in text-to-image (T2I) synthesis. While these generative AI models showcase remarkable creativity, their unrestricted nature may not consistently align with practical applications, particularly in biomimetic design.

Biomimicry involves replicating natural structures and processes to address complex challenges (Vincent et al., 2006). Meanwhile, the design space must be bounded by factors such as living structures, movement mechanisms, and scientific principles. To address this challenge, we propose a framework for Controllable Generative AI in Biomimetic Design.

Our proposed framework utilizes Large Language Models (LLMs) and morphology-constrained text-to-image (T2I) transformation. It begins with the generation of a table of entities associated with specific adjectives describing the target's nature. These entities are then projected into a unified vector space, with outliers removed based on similarity. Subsequently, adjectives are collected, ranked, and filtered to eliminate conflicting attributes. An LLM selects the best-aligned entity, and its morphological features guide the T2I transformation. This approach ensures that the generated designs align with functional requirements and real-world constraints, making them suitable for practical applications in biomimicry and related fields. By bridging the gap between AI's boundless creativity and the necessity for implementable solutions, our framework, integrating ControlNet (Zhang et al., 2023) and LLMs, holds the potential to significantly advance the realm of biomimetic design, ushering in new possibilities for innovative and sustainable solutions inspired by nature.

2. Literature Review

2.1. Stable Diffusion model (Rombach et al., 2022)

The Diffusion model differs from the other generation methods in that it needs to be run in multiple steps in terms of timing and the hidden variable z has the same dimension as the original size of x. For Stable diffusion model. The techniques involved in Stable Diffusion are Diffusion Model (DDPM), Attention, Autoencoder. It based on Latent Diffusion Models (LDMs)

Compared to the original diffusion, the advantage is that the encoder is used to map the image from the original pixel space to the latent space, which reduces the amount of computation. Since diffusion model favors semnatic compression (which is easy to be perceived by the human eye) and GAN/AutoEncoder favors perceptual compression (which means that it is difficult for the human eye to perceive the high-frequency information), the image can be similarly low-pass filtered with the encoder first. Endocer can effectively reduce the image edge length to 1/4 to 1/16 of the original, and even improve the quality of the generated samples.

2.2. Autoregressive model: CLIP and DALLE (both from OpenAI)

2.2.1. CLIP: Contrastive Language-Image Pre-training (Radford et al., 2021; Ramesh et al., 2022)

The approach of CLIP centers on leveraging a large corpus of image-text pairs to enforce alignment within a shared feature space. In this framework, a 12-layer transformer is employed for text encoding, while image encoding can utilize either ResNet or Vision Transformer (ViT), with ViT generally outperforming ResNet in evaluations.

The process is simplified to include linear projection of encoder representations to a multi-modal embedding space, omitting non-linear projection. Symmetric cross-entropy loss is employed alongside the ADAM optimizer, with training initiated from scratch. Notably, the task is defined as text-to-image (T2I) generation rather than exact word-to-image mapping in a single step, as the latter can be challenging to train. Each batch consists of N texts and N images, with the model generating N^2 scores.

Regarding data, CLIP relies on a dataset comprising approximately 400 million Text-Image pairs known as WebImage-Text, akin in scale to the WebText dataset used in GPT-2. Data augmentation is limited to random square crops from resized images.

Despite its strengths, CLIP exhibits limitations in specialized domains such as EuroSAT/MNIST, and abstract tasks like predicting the number or positional distance of items within an image, where zero-shot performance remains constrained. Overcoming these limitations may necessitate enhanced prompt engineering, sometimes involving iterative trial-and-error approaches.

2.2.2. DALLE: Language-Image Pre-training (Ramesh et al., 2021)

In the DALL-E model, images undergo encoding via the VQGAN encoder, translating them into a sequence of tokens. Simultaneously, language descriptions are encoded using the BART encoder. These encoded descriptions and images are then fed into the BART decoder, an autoregressive model that predicts the next image. The loss function employed is softmax cross-entropy, computed between the model's prediction logits and the actual image encodings derived from the VQGAN.

During the inference stage, the caption is encoded by the BART encoder, and the beginning-of-sequence token is input to the BART decoder. Image tokens are sequentially sampled based on the decoder's predicted distribution over the next token, and sequences of image tokens are subsequently decoded by the VQGAN decoder. CLIP is then utilized to select the best-generated images from the output..

2.3. Kandinsky T2I Synthesis with Image Prior and Latent Diffusion (Razzhigaev et al., 2023)

The Kandinsky model represents an enhanced iteration incorporating latent diffusion and image prior techniques. Notably, it integrates CLIP-image embeddings instead of standalone text encoders, improving image quality. Additionally, it adopts an image-prior approach, leveraging diffusion and linear mappings between CLIP's text and image embedding spaces while retaining additional conditioning with XLMR text embeddings. This dual-encoder setup encompasses CLIP-text with image prior mapping and XLMR, which remain frozen during the training phase.

A pivotal consideration driving the model's design was the efficiency observed in training latent diffusion models, as highlighted in Rombach et al. (2022). The model operates through three stages: text encoding, embedding mapping (image prior), and latent diffusion. A transformer-encoder model is employed at the embedding mapping stage, also referred to as the image prior. This component is trained from scratch using a diffusion process on text and image embeddings provided by the CLIP-ViT-L14 model. Notably, the model incorporates element-wise normalization of visual embeddings based on full-dataset statistics to expedite the convergence of the diffusion process. In the inference stage, inverse normalization is utilized to revert to the original CLIP-image embedding space.

3. Problem Statement

Text-to-image (T2I) synthesis, which generates visually realistic images from textual descriptions, stands at the crossroads of deep learning and computer vision, promising transformative possibilities. Recent strides in generative models like Generative Adversarial Networks (GANs), autoregressive models, and diffusion models have significantly expanded the horizons of this domain. Nevertheless, amidst these advancements, persistent challenges impede the seamless creation of high-quality, diverse, and contextually aligned images from textual prompts.

A fundamental hurdle lies in the diversity and fidelity of generative models themselves. GANs, renowned for their capacity to produce lifelike images, grapple with issues like mode collapse, training instability, and hyperparameter sensitivity (Sauer & Geiger, 2021; Li et al., 2023). Achieving the delicate equilibrium between generator and discriminator networks demands meticulous parameter tuning to avert subpar outcomes . Furthermore, the adversarial dynamics of GANs often hinder precise control over generated imagery and the faithful portrayal of textual prompts (Kang et al., 2023). Autoregressive models, which construct data iteratively based on preceding information, struggle to ensure coherence and capture intricate textual nuances effectively (Esser et al., 2021). Their sequential nature can constrain their ability to grasp long-range dependencies and craft highly detailed visuals (Li et al., 2021). Moreover, as input sequences lengthen, the computational demands of autoregressive models escalate, posing scalability challenges for generating high-resolution images (Du et al., 2022). While diffusion models present a promising avenue, they introduce their own complexities. The iterative process of adding and reversing noise demands significant computational resources and meticulous optimization to maintain image quality throughout denoising iterations (Meng et al., 2021).

Beyond the idiosyncrasies of individual generative models, a broader challenge looms regarding these techniques' practical deployment and adaptability in real-world contexts. Integrating T2I synthesis into arenas such as image editing, gaming, and computer-aided design mandates tackling issues of scalability, interpretability, and user-centric customization (Wang et al., 2021). Ensuring generated images resonate with user preferences, exhibit consistency across diverse textual inputs, and seamlessly integrate into existing workflows poses an ongoing dilemma (Mayerson, 2023).

Addressing these challenges and propelling the T2I synthesis field forward necessitates comprehensive inquiries into their implications on generative model efficacy. Regrettably, comparative studies evaluating model performance across various challenges remain scarce in the current literature (Zhou & Shimada, 2023). While modern AI generators dazzle with their creativity, their unbounded imaginings often diverge from practical feasibility. In contexts like biomimicry or practical applications such as housing construction or biotechnology, unrestrained creativity risks impracticality. Constraints like structural integrity, biomechanics, airflow, and scientific principles demand a harmonious fusion of creativity with functionality. Effective creativity often hinges on seamlessly blending reality with imagination.

Thus, we present a framework to bridge the chasm between AI-generated creations and real-world applicability, particularly in biomimicry. By facilitating a synergy between AI generation and practical constraints, we aspire to unlock novel applications and visionary possibilities.

4. Methodology

4.1. Generation Across Models

In our experiment, we deploy four models for text-to-image (T2I) tasks: the Stable diffusion, DALLE$_{mini}$, CLIP, and Kandinsky models. Our approach entails generating images corresponding to a series of input sentences, each crafted to present a different level of complexity. These sentences range from simple depictions of everyday scenes to more intricate and abstract descriptions. Starting with straightforward imagery, the sentences gradually introduce nuances that challenge the model's ability to interpret and visualize the intended scenes. From descriptions of nature to celestial phenomena, the complexity of the vocabulary and concepts increases, pushing the boundaries of the model's understanding and creativity. Additionally, some sentences incorporate poetic language and unconventional scenarios, adding layers of complexity that require imaginative interpretation. Through this diverse set of input sentences, the method aims to explore the model's capacity to generate visually compelling images across a spectrum of linguistic complexity and conceptual abstraction. Furthermore, to extend the capabilities of controllable generative models, we deliberately select two sentences and compare the ControlNet's output with those from the four synthesis models, providing a comprehensive evaluation of performance with controllable constraints.

4.2. Metric and Configurations

Evaluation of T2I model can be different. While, most of the machine evaluation process should be included in the training and testing process of the model which requires great load of memory and strong GPU. Here we only perform Human evaluation and IS Analysis to scrutinize the effectiveness of the models.

4.2.1. Human Evaluation

Human Evaluation focuses on confirming the credibility and alignment of automatic evaluations through user analysis. In our experiment, 8 students were recruited to conduct the evaluation. Each Participant is asked to rate generated images based on plausibility, including factors such as object accuracy, counting, positional alignment, or image-text alignment, as well as naturalness, assessing whether the image appears realistic. Ratings are provided on a 5-Point Likert scale, with 5 representing the best and 1 representing the worst. Additionally, human evaluation is emphasized for assessing challenging tasks and rare object combinations, as it ensures accurate assessments and helps avoid bias related to race or gender.

4.2.2. Inception Score

Inception Score(IS) is a frequently used metric for analyzing the performance of a generative model. A common form of IS can be expressed as follows:

$$IS(G, z) = exp\left(\mathbb{E}_{\mathbf{x}\sim P(z)} D_{KL}(p(y|\mathbf{x})||p(y))\right) \quad (1)$$

Where $D_{KL}()$ denotes the KL divergence distance between two distributions, and $p(y|x)$ and $p(y)$ denotes the conditional distribution and the marginal distribution. y is the label prediction distribution given some classification model (usually a deep neural network called the Inception V3 is used), and $\mathbf{x} \sim P_n$ are images generated from a given generative model G sampled from the set of possible outcome distribution $P(z)$ given some text input vector z. In reality we don't have the true distribution $P(z)$ nor the conditional and marginal distribution on y, hence we can only use a emperical version of it. Let N be the total number of samples generated from the model G given a input vector z, and the set $X_N(z) = \{x_1(z), \ldots, x_N(z)\}$ to be the set of generated images. Then we approximate $p(y)$ with:

$$\hat{p}(y, z) = \frac{1}{N}\sum_{i=1}^{N} p_{classifier}(y|x_i(z)) \quad (2)$$

Here $p_{classifier}(y|x_i(z))$ is the distribution output of some chosen classifier given an input image $x_i(z)$. Then the emperical version of the IS score can be written as:

$$\hat{IS}(G, z) = exp\left(\frac{1}{N}\sum D_{KL}(p_{classifier}(y|x_i(z))||\hat{p}(y, z))\right) \quad (3)$$

In our experiment, the classifier that we use is *placeholder*, and we choose N =*placeholder* for all models.

4.3. Controllable Generative AI in Biomimetic Design

While modern AI generators can inspire with their creativity, their limitless imaginings do not easily translate into reality. Were such unconstrained creations employed in biomimicry or other practical applications like housing construction or biotechnology, they would likely prove infeasible. Factors such as living structures, movement mechanisms, airflow, geometry, and science constrain expression - creative flourishes must align with the functional. And the most impactful creativity is often the one integrating reality with imagination. Hence, Controlable generative AI is a promising advance, and by incorporating greater constraints and control over the creative process, it will open up even more opportunities in AI-powered fields in the future. On this point, here we provide a demo of how the Biomimetics process can benefit from the Large Language model (LLM) and the Controllable generative AI.

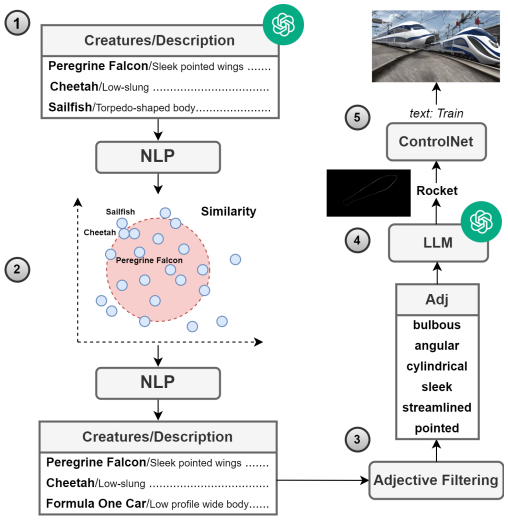


Figure 1. Framework for Integrating Controllable Generative AI into Biomimetic Design Processes.

Our framework comprises five steps, as shown in Figure 1:

- **Step1. Entity Generation and Selection**

  The initial step in our approach involves generating a table of entities related to a specific adjective that describes the nature of the target. Two methods can be employed for this purpose. The first involves scraping existing data from online encyclopedias, such as Wikidata, or acquiring datasets from platforms like Kaggle or other databases. Alternatively, large language models (LLMs) can be directly accessed to assist in this process. LLMs, such as ChatGPT, which have made recent breakthroughs in Artificial Intelligence accessible to everyone, are powerful machine-learning models capable of understanding and generating natural language. These models are trained on extensive datasets of text and code, enabling them to learn patterns and relationships within the language. Utilizing interactive platforms, LLMs can generate a table of entities

and their descriptions that align with the specified requirements. It is important to note that the descriptions can encompass various aspects of the entities' nature, with morphology being the most straightforward to visualize.

- **Step2. Entity Projection and Outlier Removal**

  A further processing step is crucial to focusing on the primary common features. This involves projecting all entities into a vector space based on their descriptions and performing dimension reduction to align them in a unified space. Subsequently, outliers exhibiting large discrepancies are identified and removed based on their similarity using Natural Language Processing (NLP) methods. In our implementation, we utilize the open-source *Spacy* library for our NLP tasks.

- **Step3. Adjective Collection and Filtering**

  Following the morphological adjective similarity comparison, a bag of entities is obtained. All adjectives associated with these entities are collected and ranked based on their frequency of appearance. An additional step involves filtering out conflicting adjectives, such as "small" and "big" or "tall" and "short." If conflicting adjectives are present, both should be removed as they do not represent the target's primary influential nature.

- **Step4. Entity Selection using LLMs**

  Utilizing the refined list of adjectives, an LLM is employed to select an entity that best aligns with all the adjectives. This process aims to identify an existing entity in the natural world that closely matches the desired characteristics. An online figure that best represents the target is then selected.

- **Step5. Morphology-Constrained T2I Generation**

  Image analysis techniques are applied to the selected figure to extract its features. These features are then provided to the ControlNet as constraints to guide the T2I transformation process. By incorporating the constraints derived from the selected entity's morphological features, the generated images are expected to exhibit a fusion of the target image characteristics and the constraining features.

5. Experimental Results

In this section, we compare the generative results of four different methods: Stable Diffusion, DALL-E, CLIP, and Kandinsky. We will present some generated results of each method based on given text sentences in Section 5.1. Following that, we conduct qualitative and quantitative comparisons across these models in Section 5.2. Subsequently, we expand the constraints with ControlNet in Section 5.3. The implementation details of the framework for Biomimetic design process integrating controllable generative AI and LLMs are provided in Appendix A.

5.1. Results Across Different Models

Input Sentences: The sentences were designed with varying levels of difficulty:

- *A red apple sits on a green leaf.* This sentence uses simple vocabulary, which is easy to understand because it depicts scenes from daily life.
- *An old oak tree stands tall in the forest, its branches reaching toward the sky.* Things get a little more complicated when you add descriptions that include adjectives like "old" and "tall." However, the concept of trees in the forest is still easy to imagine.
- *The fantastic glow of celestial bodies illuminates the canvas of the night, creating a mesmerizing tapestry of cosmic beauty.* This term uses more sophisticated vocabulary such as "celestial," "mesmerizing," and "tapestry." Thus, the resulting images also become more abstract and complex.
- *A firefly snakes its way through the velvety darkness of the forest.* More descriptive vocabulary, such as "snakes its way" and "velvety darkness." The model needs to generate a more poetic scene, rooted in nature.
- *The cat rides the horse.* Although simple vocabulary is used here, the concept of a cat riding a horse is so bizarre that it is difficult to deduce from modern image databases, ie., absurdity increases complexity.

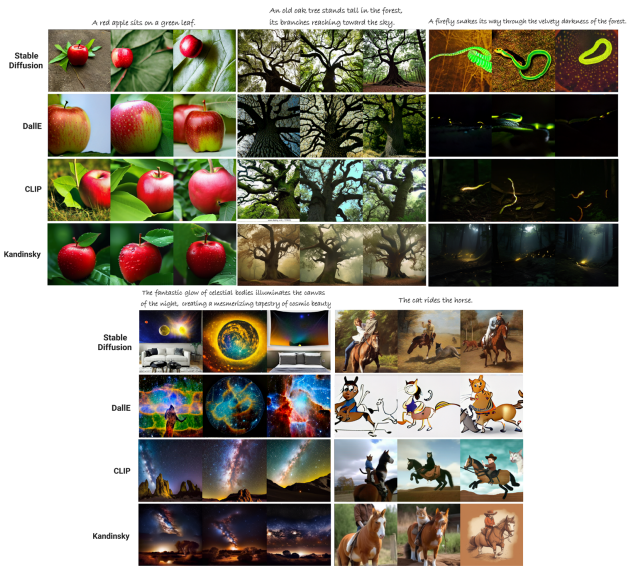


Figure 2. Generative Results across different models.

5.2. Evaluation

5.2.1. Inception Score

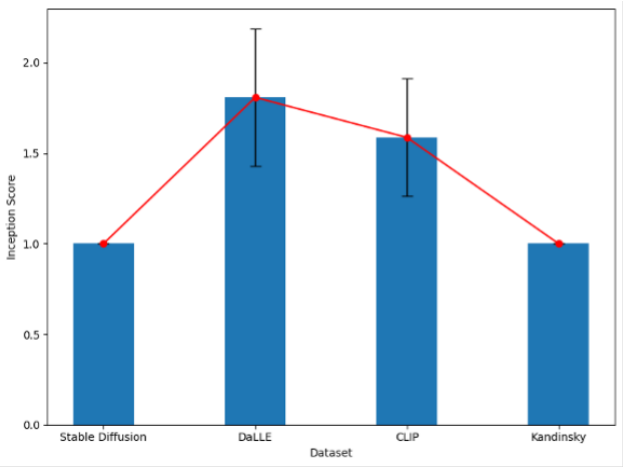


Figure 3. The empirical IS Score generated from the 4 models.

Based on Figure 3, Stable Diffusion exhibits the lowest IS score. The IS score serves as a measure of a generative model's consistency in producing content. For instance, if the input text is "*the cat rides the horse*," $p_{\text{classifier}}(y|x_i(z))$ represents the prediction of a classifier using a generated image $x_i(z)$, while $\hat{p}(y, z)$ signifies the average distribution of all images generated from the model based on the input text "*the cat rides the horse*." Although these methods aim to offer a variety of images for users to choose from, consistency in content is still desired.

A small average KL distance indicates that all generated images from the model closely resemble the classifier's predictions. If the classifier is appropriately selected, it suggests that all images from the generative model belong to the same class as indicated by the classifier. Based on this understanding, we observe that both Stable Diffusion and Kandinsky's method notably outperform the remaining two methods.

5.2.2. Human Evaluation

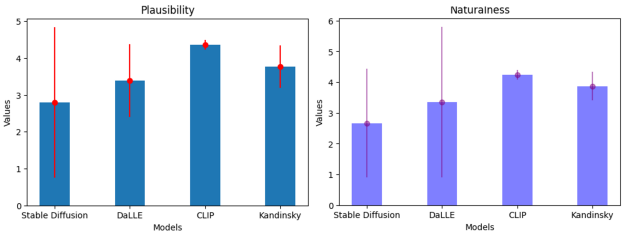


Figure 4. Plausibility and Naturalness Evaluation.

The human evaluation revealed CLIP generated the most plausible images compared to the other models. Survey results showed most annotators felt CLIP and Kandinsky produced more sensible outputs than the rest. Although DaLLE achieved the highest IS score - indicating greater diversity - participants found its artistic style less realistic. Still, having more deviation bolstered diversity, playing to DaLLE's strength. In terms of naturalness though, DaLLE underperformed other models due to its creative stylization."

5.3. ControlNet

Control network is an extension of diffusion model, which shows a promising new method of adding spatial control to the process of T2I conversion by allowing additional conditional inputs (such as segmentation maps, depth maps, human bodies, gesture key points, etc.). Under the backbone of diffusion network, it provides a more accurate control direction in the stable diffusion backbone network than only obtaining clues from text. This allows AI to produce images that are more in line with our visual imagination, rather than merely fusions of features. A key innovation of ControlNet is its method of fine-tuning powerful reservation functions, which spread steadily in an end-to-end way without causing catastrophic forgetting or quality degradation. This makes it possible to learn specialized conditional tasks with smaller datasets. Collecting labeled elements and data is always expensive. Therefore, ControlNet limits the infinity of characteristics from the representative features of a small amount of data, giving us a great opportunity to get the desired results.

To assess the effectiveness of ControlNet, we conducted two tests on our sentence generation tasks:

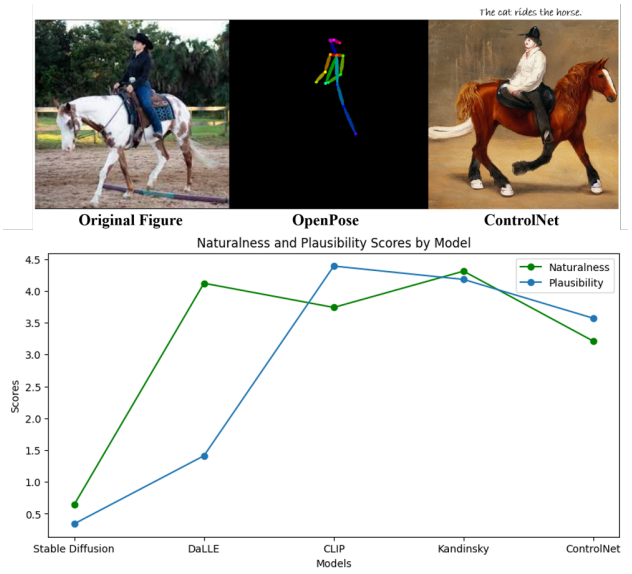


Figure 5. T2I Generation with Pose-strengthened ControlNet.

5.3.1. THE CAT RIDES THE HORSE

For this task, we selected a relevant image of a cowgirl online and utilized the OpenPose toolkit to extract key body points representing her posture. These points were then input into ControlNet to constrain the movements of the main character described in the input text, in this case, the cat. As a result, 'riding' confined the cat's gesture, reflecting the expected posture. Evaluation of the images generated by ControlNet and other models revealed a significant improvement over the original Stable Diffusion method, approaching the quality of CLIP and Kandinsky. Human annotators consistently rated the naturalness and plausibility of ControlNet-generated images higher.

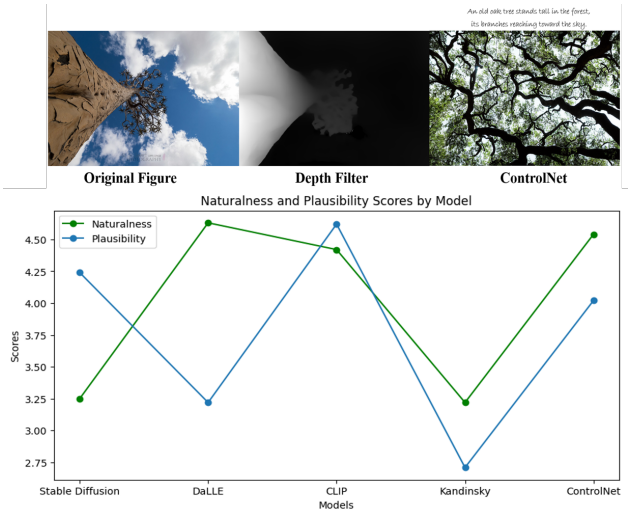


Figure 6. T2I Generation with Pose-strengthened ControlNet.

5.3.2. AN OLD OAK TREE STANDS TALL IN THE FOREST, ITS BRANCHES REACHING TOWARD THE SKY

In this task, the phrase "reaching toward the sky" describes the dynamic vision of the disparity between the tree's branches and its trunk, highlighting the need to represent the tree's depth accurately. Based on the provided depth filter, ControlNet adeptly captures the depth and reach of tree branches through spatial and style constraints, resulting in highly realistic renderings. Comparative assessment against DaLLE and the original Diffusion model confirmed ControlNet's superior quality, as determined through qualitative human assessment. These results suggest that ControlNet successfully specifies the necessary spatial and style constraints through fine-tuning specific images. These experiments underscore ControlNet's robust performance in enhancing image generation tasks, contributing to generative models' advancements.

Conclusion and Future Direction

In this study, we comprehensively evaluated four state-of-the-art text-to-image synthesis models: Stable Diffusion, DALL-E, CLIP, and Kandinsky. Our qualitative and quantitative comparisons revealed each model's strengths and limitations in generating high-quality, diverse, and semantically aligned images from textual descriptions. The results highlighted the superior performance of Controllable Generative AI, particularly when integrated with Large Language Models (LLMs) for enhanced biomimetic design processes. Future research should refine the proposed framework for Controllable Generative AI in Biomimetic Design, exploring novel techniques to incorporate real-world constraints and domain-specific knowledge into the creative process. Additionally, investigating the scalability and adaptability of these models to various application domains, such as architecture, product design, and biotechnology, will be crucial for unlocking their full potential. By bridging the gap between AI-generated creativity and practical feasibility, we can pave the way for innovative and sustainable solutions inspired by nature's wisdom.

References

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41 (2):423–443, 2018.

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A. W., Firat, O., et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pp. 5547–5569. PMLR, 2022.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Kang, M., Zhu, J.-Y., Zhang, R., Park, J., Shechtman, E., Paris, S., and Park, T. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10124–10134, 2023.

Li, B. Z., Nye, M., and Andreas, J. Implicit representations of meaning in neural language models. *arXiv preprint arXiv:2106.00737*, 2021.

Li, R., Li, W., Yang, Y., Wei, H., Jiang, J., and Bai, Q. Swinv2-imagen: Hierarchical vision transformer diffusion models for text-to-image generation. *Neural Computing and Applications*, pp. 1–16, 2023.

Liu, A. H., Jin, S., Lai, C.-I. J., Rouditchenko, A., Oliva, A., and Glass, J. Cross-modal discrete representation learning. *arXiv preprint arXiv:2106.05438*, 2021.

Mayerson, D. R. Gravitational multipoles in general stationary spacetimes. *SciPost Physics*, 15(4):154, 2023.

Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Razzhigaev, A., Shakhmatov, A., Maltseva, A., Arkhipkin, V., Pavlov, I., Ryabov, I., Kuts, A., Panchenko, A., Kuznetsov, A., and Dimitrov, D. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Sauer, A. and Geiger, A. Counterfactual generative networks. *arXiv preprint arXiv:2101.06046*, 2021.

Van Looveren, A., Klaise, J., Vacanti, G., and Cobb, O. Conditional generative models for counterfactual explanations. *arXiv preprint arXiv:2101.10123*, 2021.

Vincent, J. F., Bogatyreva, O. A., Bogatyrev, N. R., Bowyer, A., and Pahl, A.-K. Biomimetics: its practice and theory. *Journal of the Royal Society Interface*, 3(9):471–482, 2006.

Wang, Y., Qi, L., Chen, Y.-C., Zhang, X., and Jia, J. Image synthesis via semantic composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13749–13758, 2021.

Yadav, A. and Vishwakarma, D. K. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385, 2020.

Zameshina, M., Teytaud, O., and Najman, L. Diverse diffusion: Enhancing image diversity in text-to-image generation. *arXiv preprint arXiv:2310.12583*, 2023.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Zhou, Y. and Shimada, N. Vision+ language applications: A survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 826–842, 2023.

A. Appendix

A.1. Biomimetic Design: High-Speed Train

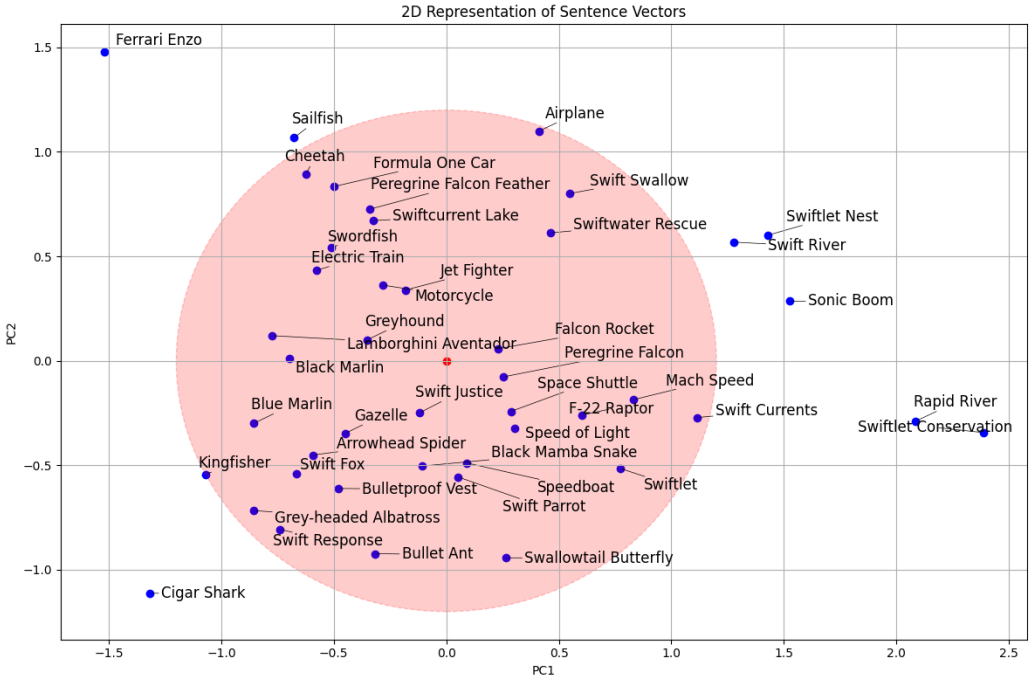
The generative models discussed previously have showcased remarkable creativity, enhancing the artistic domain of Biomorphism and displaying versatility across various creative avenues. However, creativity alone may not suffice to meet the demands of Biomimeticists. Drawing inspiration from nature's intricate designs and functionalities, biomedical products often encounter challenges in transitioning from concept to reality. While Biomimetic scientists leverage their understanding of nature to design and meet requirements, sourcing creatures with desired abilities can pose difficulties. While our previous experiments highlighted ControlNet's potential to balance creativity and reliability, they only showcased the model's generative capabilities. Here, we propose a framework that integrates the generalization ability of Large Language Models (LLMs) with ControlNet to replicate a successful innovation, the High-Speed Train.

In this scenario, following the pipeline in Figure 1, the design process for the "High-Speed Train" unfolds as follows:

- **Entity Generation and Selection:** Generate a table of entities related to fast.

	Creature/Object	Description
0	Peregrine Falcon	Sleek pointed wings, Streamlined body long tai...
1	Cheetah	Low-slung body slender limbs, Muscular shoulde...
2	Sailfish	Torpedo-shaped body pointed bill, Muscular for...

- **Entity Projection and Outlier Removal:** Utilize Natural Language Processing (NLP) methods to identify and remove outliers based on adjective similarity.



- **Adjective Collection and Filtering:** Compile all existing adjective terms and filter out those related to shape, particularly addressing conflicting adjectives such as 'wide' and 'narrow,' 'tall' and 'short.'

- **Entity Selection using LLMs:** Leveraging OpenAI, we identify two entities: "Rocket" and "Aerodynamic Spacecraft." We acquire an image of a rocket online and convert it into a canny image.

Step 1. Pick up adjectives related to shape

```
'bulbous', 'angular', 'cylindrical', 'sleek', 'streamlined',
'pointed', 'lean', 'narrow', 'wide', 'deep', 'large',
'short', 'compact', 'broad', 'tall', 'small'
```

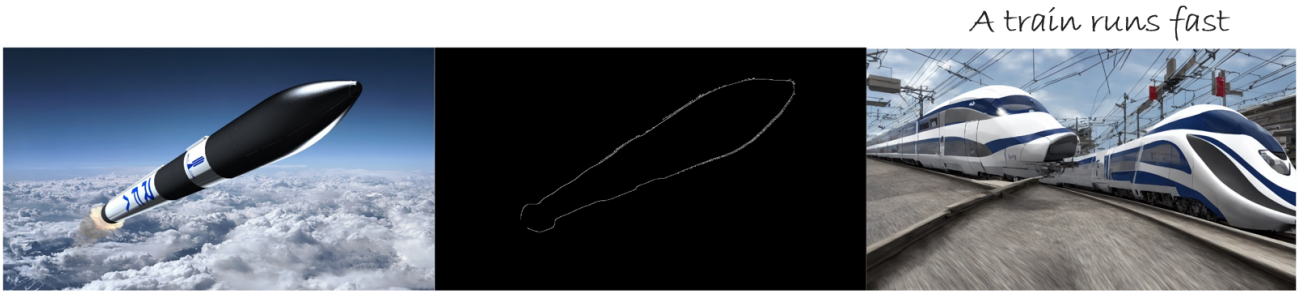
Step 2. Remove conflicting adjectives

```
'bulbous', 'angular', 'cylindrical', 'sleek', 'streamlined', 'pointed', 'lean', 'compact'
```

Step 3. Find objects that best fits the remaining adjectives:

```
'Rocket', 'Aerodynamic Spacecraft'
```

- **Morphology-Constrained T2I Transformation:** With the constraint of the rocket's contour, the ControlNet generates a promising output for the text "*A train runs fast.*"



Rocket **Rocket's Contour** **ControlNet**

Based on the generated result of the ControlNet, it's evident that the contour of the rocket effectively shapes the train, giving it a sleeker, more streamlined appearance with a pointed head. This design adjustment aids in reducing air resistance. Opting for the contour of the rocket with canny edges, rather than emphasizing depth or adding intricate details to the rocket's edges, may not sufficiently constrain the generation process. Our goal is to focus on the shape of the train rather than the patterns on it. Providing excessive detail for the constraint could potentially mislead the ControlNet's generation process, resulting in inaccurate feature fusion.